CDF-MATCHING FOR AUTOMATIC TONE ERROR DETECTION IN MANDARIN CALL SYSTEM

Si Wei, Hai-Kun Wang, Qing-Sheng Liu, Ren-Hua Wang

iFlytek Speech Lab, University of Science and Technology of China, Hefei
{tonyw, hkwang, ustc_qsliu}@ustc.edu, rhw@ustc.edu.cn

ABSTRACT

This paper introduces a tone pronunciation error detection algorithm for Mandarin CALL system. HMM is introduced to build tone model. F0 after CDF-matching normalization is used as the feature of tone model. Tone error detection algorithm is based on the posterior probability calculated from HMM tone models. Comparing to the other normalization methods, CDF-matching can get better tone error detection performance with CC's increasing from 0.76 to 0.79, which is close to that between human evaluators (0.83).

Index Terms **—** CDF-matching, Tone recognition, Pronunciation error detection, CALL

1. INTRODUCTION

Computer Assisted Language Learning (CALL) systems can provide many potential benefits for both the language learner and teacher [1]. Mandarin is a tonal language. Mandarin CALL system should be able to automatically detect the tone pronunciation error made by students and give them corrective advices. Many CALL systems investigate pronunciation error detection [3, 4]. The aim of this paper is to get an efficient tone error detection algorithm under the framework of Hidden Markov Model (HMM) for Mandarin CALL system, which is constructed to teach Chinese with dialect accent to learn Mandarin.

Previous work mainly focuses on segmental pronunciation error detection such as the work done by Witt [3]. Witt's algorithm utilizes posterior probability calculated from automatic speech recognizer for pronunciation error detection. Ito [4] introduces a pronunciation error detection algorithm with multi thresholds based on decision tree. All these methods can efficiently detect segmental pronunciation error. At the same time, many researches have been done for Mandarin tone recognition. Zhang [5] introduces tone-nuclei method to model tone. Lin [6] utilizes high level information got from f0 contour to model tone. Zhou [7] uses sub linear regression coefficients of f0 contour as the feature for tone recognition. Although there are many algorithms about tone recognition for Mandarin, little has been done for Mandarin tone error detection till now. Most of the tone recognition algorithm uses HMM to capture the movement of f0 contour. This paper also uses HMM for tone modeling and f0 as the feature for tone recognition. Then posterior probability got from the tone recognizer is utilized as the measure of tone pronunciation accuracy. If the posterior probability is below a preset threshold, a tone pronunciation error is pointed out.

F0 contour differs greatly with speaker and the style of pronunciation. Many methods are used to do normalization, such as the normalization by logarithm of f0 [8], mean normalization, mean and variance normalization, etc. The purpose of normalization is to make different person's f0 distributions to be similar, in other words, to make different cumulative distribution functions to be similar. In this paper, CDF-matching is introduced to normalize the f0 distribution by mapping different speaker's cumulative distribution function of f0 into one standard speaker's distribution.

Cross-correlation (CC) is taken as the measurement of error detection performance. CC used here is the same as defined by Witt [3], which is shown in equation (5). Based on a 60 persons' database recorded from PSC test (A national test for spoken Mandarin proficiency), CC's improvement from 0.76 to 0.79 is gained by CDF-matching normalization comparing to the other normalization methods.

This paper is organized as follows. Section two introduces the characteristic of Mandarin tone. In section three, various normalization methods including mean normalization, mean and variance normalization and CDF-matching normalization are investigated. Tone error detection algorithm is introduced in section four. After a brief view of our database, experiments based on the tone error detection algorithm with mean and CDF-matching normalization are carried out and results are given in section five. Section six gives the conclusion and the direction of future work.

2. CHARACTERISTIC OF MANDARIN TONE

This section introduces representation of Mandarin tone and tone distribution of different person.

2.1. Representation of Mandarin Tone

Mandarin is a tonal language. Tone is very important to distinguish Mandarin syllable. Mandarin tone can be represented by f0 contour [9]. which is shown in figure 1 [7].

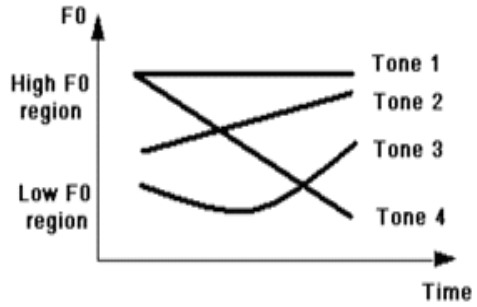

Fig. 1. Illustration of Mandarin tone types VS. f0

It seems to be easy to distinguish tone via f0 contour. But f0 contour's distinguishability is heavily influenced by variation of speaker and style of pronunciation. Compensation for the variation is urgently needed.

2.2. F0 Distribution of Different Person

F0 distributions of two different persons are shown in Figure 2, which is shown by histogram with a resolution of 1 Hz. Difference between the two f0 distributions is so large, which means that normalization for f0 is very essential for tone recognition.

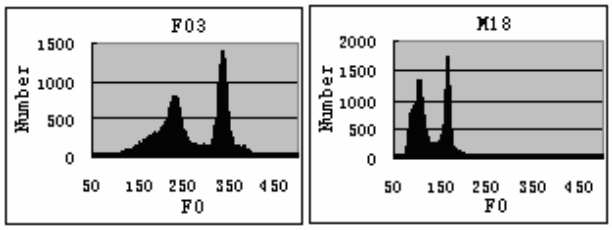

Fig. 2. F0 distributions of different people F03 and M18
The f0 data is extracted from isolated syllable of Standard Mandarin Database

3. F0 NORMALIZATION METHODS

This section introduces the methods for f0 normalization.

3.1. Mean and Variance Normalization

Mean normalization is implemented via equation (1).

$$f' = f - \overline{f} \ . \tag{1}$$

Where f and f' is the f0 value before and after normalization and $\overline{f}$ is the average f0 value of the person to be normalized.

Mean and variance normalization is implemented by equation (2).

$$f' = \frac{f - \overline{f}}{\sigma_f} \ . \tag{2}$$

Where σ_f is the variance of f0 value from the person to be normalized.

3.2. Principle of CDF-Matching

Suppose the parameter transform to be $x = T[y]$. Y is the parameter before normalization. X is the parameter after normalization. Suppose the cumulative distribution function of X is $C_X(x)$ and the cumulative distribution function of Y is $C_Y(y)$. $x = T[y]$ should satisfy $C_Y(y) = C_X(x)$. Then we can get $x = C_X^{-1}(C_Y(y))$. That means CDF-matching can be implemented by equation (3) [10].

$$x = T[y] = C_X^{-1}(C_Y(y)) \ . \tag{3}$$

3.3. Pitch Normalization Using CDF-Matching

Preceding section introduces the principle of CDF-matching. This section shows how to normalize pitch by CDF-matching.
CDF-matching is implemented as follows. Firstly a standard person F03's f0 cumulative distribution function is selected as the objective CDF function. Then all the other person's f0 distributions are mapped into F03's f0 distribution via CDF-matching. The mapping procedure is fulfilled by using histogram equalization, which is illustrated in figure 3.

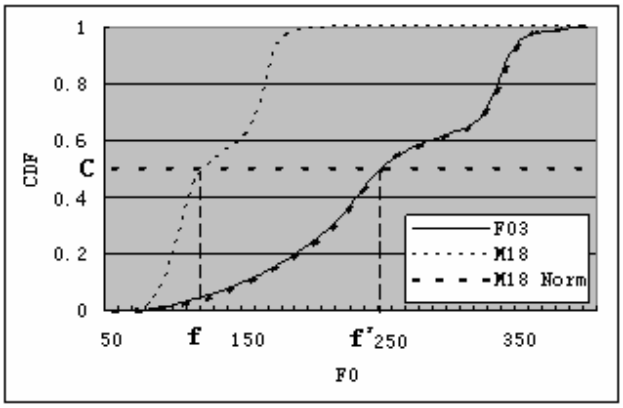

Fig.3 Mapping process of CDF-matching

The dotted lines are the CDFs of M18 before and after normalization and the solid line is of F03. CDFs are calculated by histogram. Suppose there is a frequency from

M18, which is f in figure 3. Firstly CDF value of f is calculated from the CDF of M18, which is C. Then by knowing value C, the corresponding frequency of F03's CDF is calculated, which is f' in the figure. Then f from M18 is mapped into f' by CDF-matching. This is the process of CDF-matching. It's shown in the figure that after normalization, the two CDFs of F03 and M18 become very similar.

Figure 4 gives the distribution after normalization with CDF-matching. Contrast to figure 2, f0 distribution of different people becomes very similar after CDF-matching.

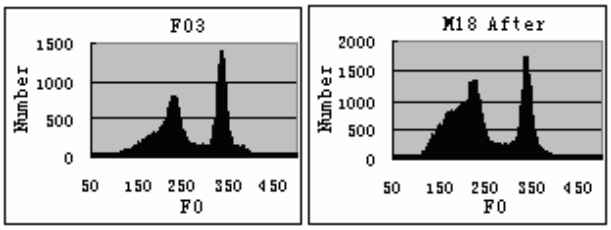

Fig. 4. Comparison between f0 distributions of M18 after normalization and standard f0 distribution of F03.

4. TONE ERROR DETECTION METHOD

Most of the algorithms use posterior probability as evaluation or error detection method. Posterior probability is used for tone error detection too. The posterior probability of tone is calculated as equation (4).

$$P(T \mid O) = \frac{P(O \mid T)P(T)}{\sum_{T' \in T_{set}} P(O \mid T')P(T')} \quad . \tag{4}$$

O is the f0 contour of one syllable. T is the tone label of text. T_{set} is the set of tone models.

HMM model with 4 emission states is used to represent the movement of f0 contour. The model is constructed by HTK. HMM tone model is used to calculate $P(O \mid T)$ and the appearance probability is supposed to be equal. The features of tone model are f0 and its first and second order derivative, which is calculated by ETSI front-end [11].

The posterior probability calculated by equation (2) is for isolated tone. For continuous tone, alignment is done by cepstrum first and then the segmentation is used to calculate tone posterior probability.

After the posterior probability is got, tone error detection is done as equation (5).

$$\begin{cases} Right & if\ P(T \mid O) \geq Thresh \\ Error & if\ P(T \mid O) < Thresh \end{cases} \quad . \tag{5}$$

$Thresh$ is the tone error detection threshold.

Equation (3) indicates that the pronounced tone is error if its posterior probability is less than threshold and right if not.

5. EXPERIMENTS BASED ON TONE ERROR DETECTION ALGORITHM

All kinds of f0 normalization method including mean normalization, mean and variance normalization and CDF-matching normalization and the algorithm to detect tone pronunciation error are introduced in preceding section. In this section, experiments using the normalization methods and the error detection algorithm on PSC test database are carried out to validate their performance.

5.1. Database and Performance Measure

Database used in this section is recorded from PSC test, which is a national test to evaluate the proficiency of spoken Mandarin. It contains four parts of test. Our tone error detection experiment is carried out on the first part of PSC test, which is reading 100 single syllable words. All of the data is used to estimate histogram. The database contains 60 persons from AnHui of China. Totally, the database contains 6000 syllables. Two professional evaluators from PSC test are invited to detect the tone pronunciation error in the database. The errors detected by human evaluator are taken as the reference for machine tone error detection algorithm. The performance of error detection is measured by the Cross-Correlation (CC) between two detection results. Cross-Correlation takes into account only those syllables where exists a rejection in either of the two judgements. The Cross-Correlation is calculated as equation (6).

$$CC_{d1,d2} = \frac{x_{d1}^T x_{d2}}{\| x_{d1} \|_E \| x_{d2} \|_E} \quad . \tag{6}$$

Here, x_{d1} is the judgement vector from evaluator d1, x_{d2} is from evaluator d2. The elements of judgement vector are 0 or 1 where 0 means tone pronunciation is right and 1 for wrong pronunciation. If both of the judgement for a segment is right, it's discarded from judgement vector in order to put emphasis on error pronunciation. $\| x \|_E = \sqrt{\sum_{i=0}^{N-1} x^2(i)}$ is the standard Euclidean distance. CC measures the similarity of rejection between two judgements. Because tone pronunciation error detected by human evaluator is the objective of tone error detection algorithm, CC is very suitable to validate the performance of tone error detection algorithm.

5.2. Performance of Human Tone Error Detection

CC between two human evaluators is listed in table 1.

Table 1. CC between two human evaluators

CC	Tone1	Tone2	Tone3	Tone4	Average
H VS H	0.86	0.82	0.78	0.86	0.83

These CCs between human evaluators are the reference for machine tone error detection algorithm and CCs between human and machine are the measure of performance about the tone error detection algorithm.

5.3. Performance of Tone Error Detection Algorithm

The tables below give the CC between tone error detection algorithms and human evaluator, which include experiments without normalization, with mean normalization, with mean and variance normalization and with CDF-matching normalization.

Table 2. CC between human and machine without normalization

CC	Tone1	Tone2	Tone3	Tone4	Average
M VS H	0.48	0.46	0.45	0.66	0.51

Table 3. CC between human and machine with mean normalization

CC	Tone1	Tone2	Tone3	Tone4	Average
M VS H	0.84	0.75	0.57	0.79	0.74

Table 4. CC between human and machine with mean and variance normalization

CC	Tone1	Tone2	Tone3	Tone4	Average
M VS H	0.85	0.74	0.67	0.78	0.76

Table 5. CC between human and machine with CDF-matching normalization

CC	Tone1	Tone2	Tone3	Tonr4	Average
M VS H	0.81	0.72	0.80	0.81	0.79

Table 2 gives the CC between human evaluator and tone error detection algorithm without normalization. The average CC (0.51) is rather low comparing to CC between humans (0.83). By using mean normalization and mean variance normalization, CC between human and machine improves to 0.74 and 0.76, which is shown in table 3 and 4. After normalization with CDF-matching, average CC between human and machine (0.79) increases further. Table 5 gives the result. The results indicate that normalization with CDF-matching is better than mean normalization and mean variance normalization. The reason is that CDF-matching can shift the f0 distribution more precisely to the standard distribution than the others methods. But CDF-matching needs more data to get robust distribution estimation than the other methods, which is unpractical in some circumstances. But if data is sufficient, just like the PSC test, CDF-matching can obtain much better performance.

6. CONCLUSION

This paper uses HMM to build tone model and introduces an algorithm to detect tone pronunciation error based on the posterior probability calculated from HMM-based tone model. At the same time, CDF-matching is used to normalize the f0 distribution. After CDF-matching normalization, the CC between human and machine tone error detection increases greatly comparing to the other normalization methods. The result indicates that the CDF-matching is useful. Comparing to the CC between humans (0.83), the CC between human and machine is close to it (0.79). But CDF-matching needs more data to robustly estimate the distribution of f0. How to estimate the f0 distribute with little data is the direction of future work. At the same time, this paper only investigates tone error detection for isolated syllables. Detecting tone pronunciation error for continuous speech is also our future work.

7. REFERENCES

[1] S. M. Witt, "Use of Speech Recognition in Computer-assisted Language Learning", PhD thesis, Cambridge, 1999

[2] H.Franco, L.Neumeyer, Y.Kim and O.Ronen, "Automatic Pronunciation Scoring for Language Instruction", *Proc. ICASSP*, pp. 1471-1474, 1997.

[3] S.M. Witt, S.J.Young, "Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning", *Speech Communication*, pp. 95-108, 2000.

[4] A. Ito, Y. Lim, M. Suzuki and S. Makino "Pronunciation Error Detection Method based on Error Rule Clustering using a Decision Tree", *Proc. EuroSpeech*, pp. 173-176, 2005.

[5] J. Zhang, K. Hirose, "Tone Nucleus Modeling for Chinese lexical Tone Recognition", *Speech Communication*, pp. 447-466, 2004.

[6] W. Lin, L. Lee, "Improved Tone Recognition for Fluent Mandarin Speech based on New Inter-syllabic Features and Robust Pitch Extraction", *Proc. ASRU*, pp. 237-242, 2003.

[7] J. Zhou, Y. Tian, Y. Shi, C. Huang and E. Chang, "Tone Articulation Modeling for Mandarin Spontaneous Speech Recognition", *Proc. ICASSP*, pp. 997-1000, 2004.

[8] M.K. Sonmez, L. Heck, M. Weintraub, and E. Shriberg. "A Lognormal Tied Mixture Model of Pitch for Prosodybased Speaker Recognition". *Proc. Eurospeech*, pp. 1391-1394, 1997.

[9] W. S. Y. Wang, "Phonological Features of Tone", International Journal of American Lingustics, pp. 93-105, 1967.

[10] J.C. Segura, "CDF-matching based Nonlinear Feature Transformations for Robust Speech Recognition", presentation, 2002.

[11] ETSI standard doc, "Extended Advanced Front-end Feature Extraction Algorithm", ETSI ES 202 050 Ver.1.1.2. 2005.